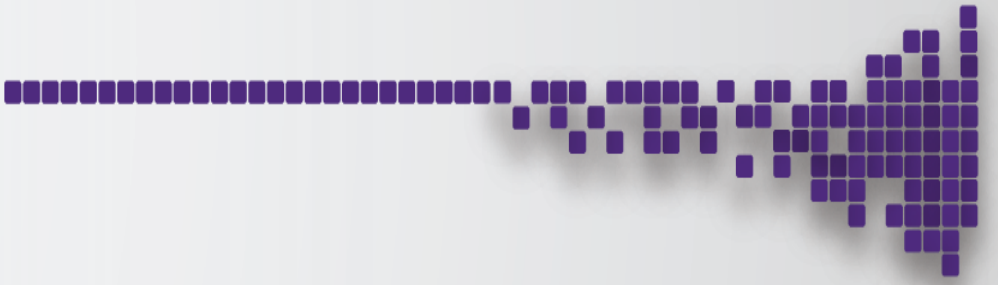
Large-Scale Data Analytics Workflow Support for Climate Change Experiments



INDIGO - DataCloud

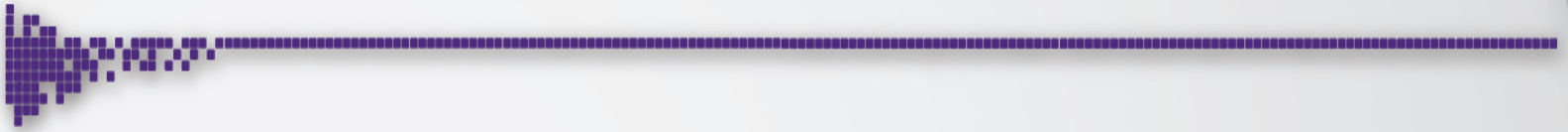
S. Fiore, C. Doutriaux, D. Palazzo, A. D'Anca, Z. Shaeen,
D. Elia, J. Boutte, V. Anantharaj, D. N. Williams, G. Aloisio

INDIGO-DataCloud is co-founded by the
Horizon 2020Framework Programme

INDIGO-DataCloud

- **An H2020 project** approved in January 2015 in the EINFRA-1-2014 call
  - 11.1M€, 30 months (**from April 2015 to September 2017**)
- **Who**: **26 European partners** in 11 European countries
  - Coordination by the Italian National Institute for Nuclear Physics (INFN)
  - Including developers of distributed software, industrial partners, research institutes, universities, e-infrastructures
- **What**: **develop an open source Cloud platform** for computing and data ("DataCloud") tailored to science.
- **For**: **multi-disciplinary scientific communities**
  - E.g. structural biology, earth science, physics, bioinformatics, cultural heritage, astrophysics, life science, climatology
- **Where**: deployable on **hybrid (public or private) Cloud infrastructures**
  - INDIGO = **IN**tegrating **D**istributed data **I**nfrastructures for **G**lobal Expl**O**itation
- **Why**: answer to the technological **needs of scientists** seeking to easily exploit distributed Cloud/Grid compute and data resources.

INDIGO & the Climate Model Intercomparison Data Analysis case study



- The proposed case study is mainly related to the climate change community
- It is directly connected to the **Coupled Model Intercomparison Project** (CMIP) and to the **Earth System Grid Federation** (ESGF) infrastructure
- A **EU/US testbed** has been setup at CMCC, LLNL, ORNL and PSNC to demonstrate the feasibility of the approach and provide real feedback to end users
- Preliminary results have been presented by Valentine G. Anantharaj (ORNL) at the **IEEE Big data 2016** conference this week
  - S. Fiore et al, "*Distributed and cloud-based multi-model analytics experiments on large volumes of climate change data in the Earth System Grid Federation eco-system*", IEEE Big Data Conference 2016, December 5-8, 2016, Washington [to appear].

The context of the case study: ESGF and the CMIP5 data archive

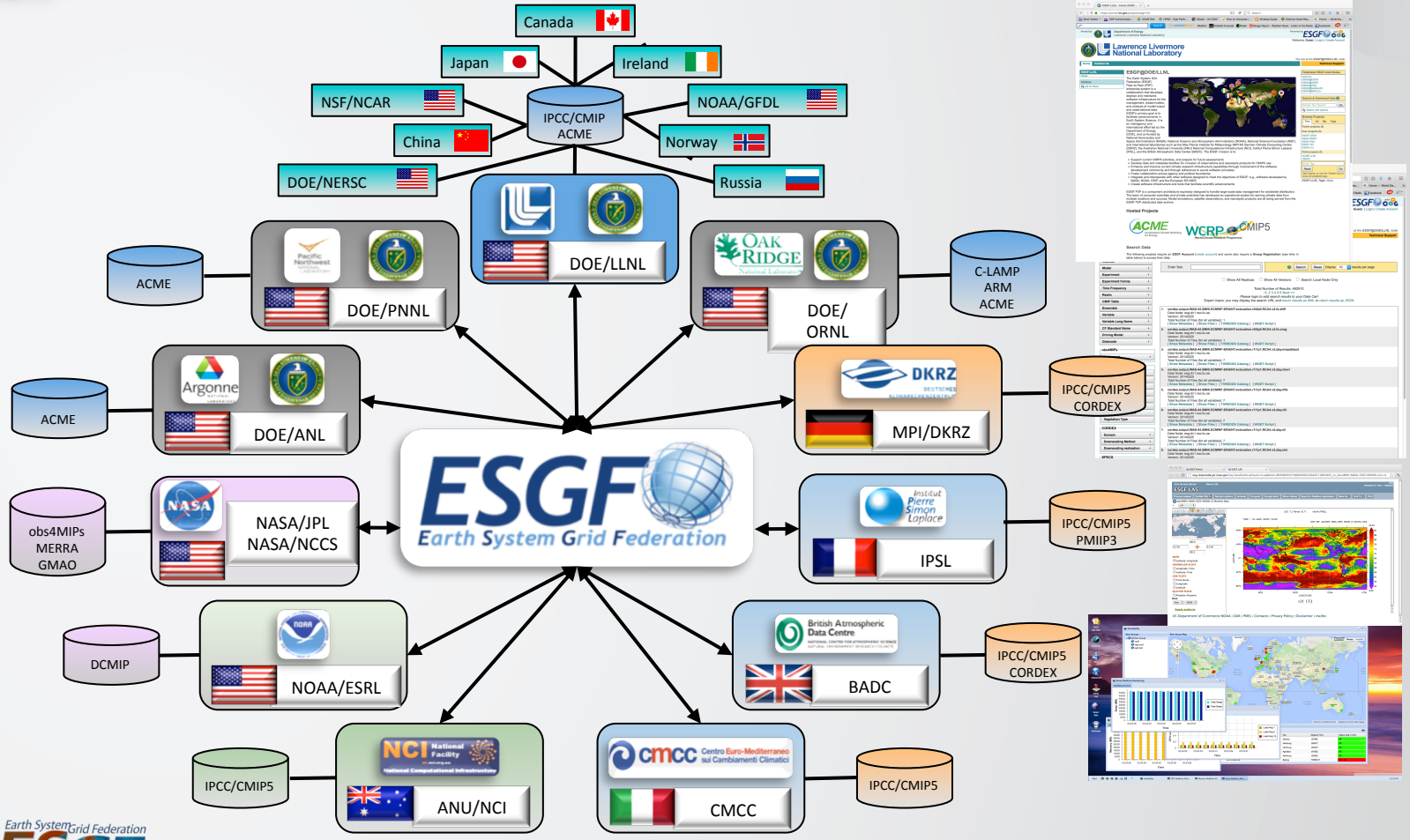


Image courtesy: Dean N. Williams (LLNL)

Requirements analysis for the climate change case study

INDIGO - DataCloud

ENES - CMCC	ENES# 1	Deployment of a software	Computing / PaaS	M	Dedicated installation of the	Easy to deploy	More flexibility, higher level of supported platforms and multi-
	ENES# 7	Isolation of deployments	Computing	C	Currently users share the	Unavailable feature	Need for minimising side-effects and Ophidia deployments are the reference data.
	ENES# 11	Metadata management.	Storage / PaaS Service	C	The system uses Thredds for managing catalogues/meta data, Solr index for indexing datasets	Available	Keep feature
	ENES# 12	Authentication and Authorization	Security / PaaS Service	M	Federated identity based on OpenID mechanism. Myproxy servers are also available.	Available for data sharing only	it should be extended to big data analysis facilities for running intercomparison experiments.
	ENES# 13	External restricted access	Security / PaaS Service	C	Anonymous access to web portals and scientific gateways	Not identified	Specific deployment with limited data analysis functionalities could serve for demo, training, dissemination.
	ENES# 14	Interactive processing	Computing / PaaS Service	C	Interactive processing is available client-side	Server-side approach should provide interactive processing capabilities	To be made available in a distributed, server-side processing/analysis scenario. Software like Ophidia and IPython deal with interactive data analysis aspects.
	ENES# 15	Easy-to-use environment	Security/Co mputing	M/ O	Set of tools for data analysis, processing,	No scientific gateways tailoring data analytics	-Data Analytics Gateways for complex experiments/workflows and high resolution data for

when exhausting
pabilities of one
t or when combining
ing of different data
e deployed on different
tics infrastructures.

easy to deploy a self-
e and auto-scalable
tics cluster with all the
ıd the console /
ser interface without
inistration
l.

of papers and
(for provenance and
y). Marketplace
r sharing workflows.
taking into account
rk in the area (e.g.

emand. Dynamicity can be
onsidered Mandatory, whereas
lasticity can be considered
ptional.

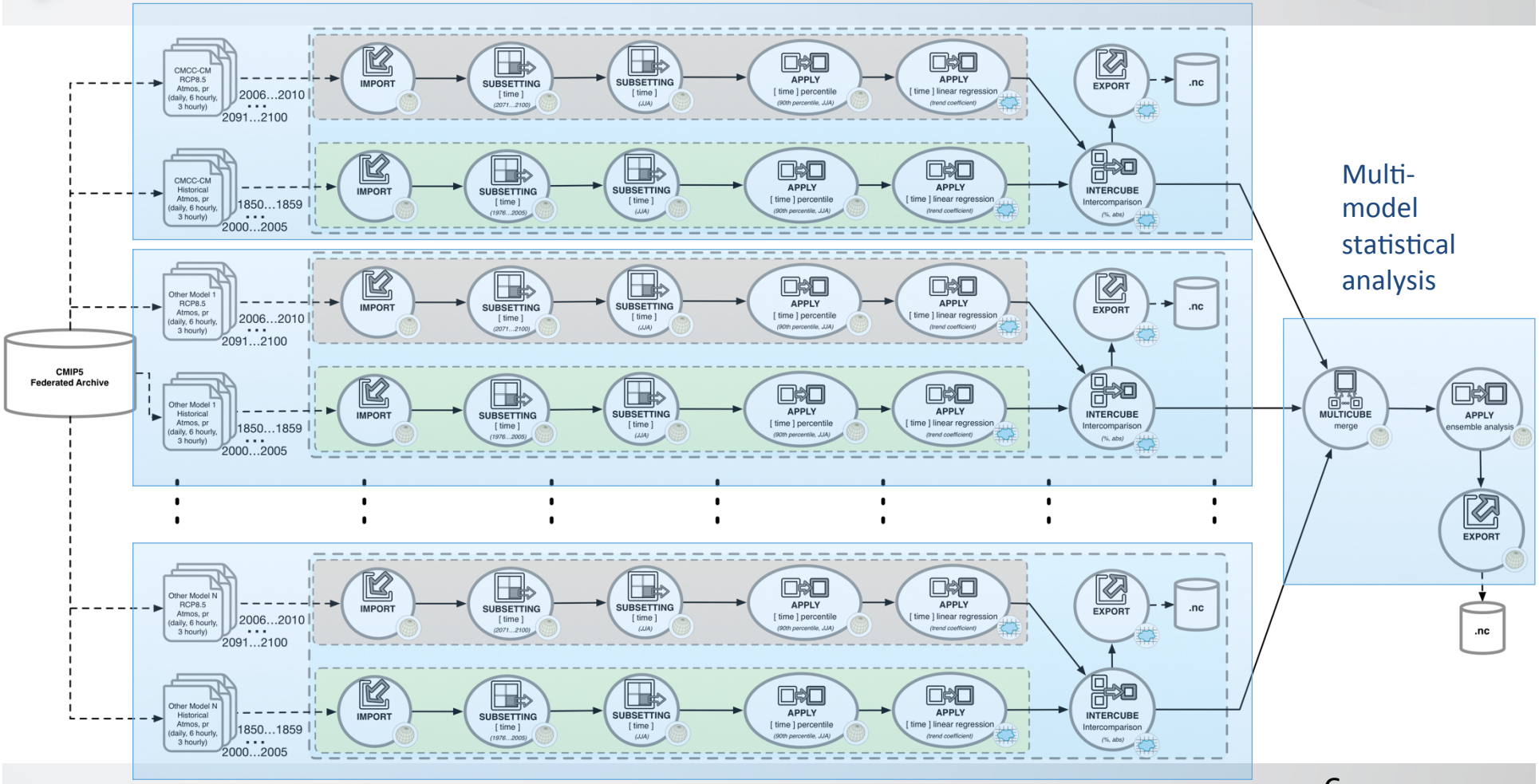
he
w" to
lysis

to
he

High-level view of the multi-model experiment on Precipitation trend analysis

INDIGO - DataCloud

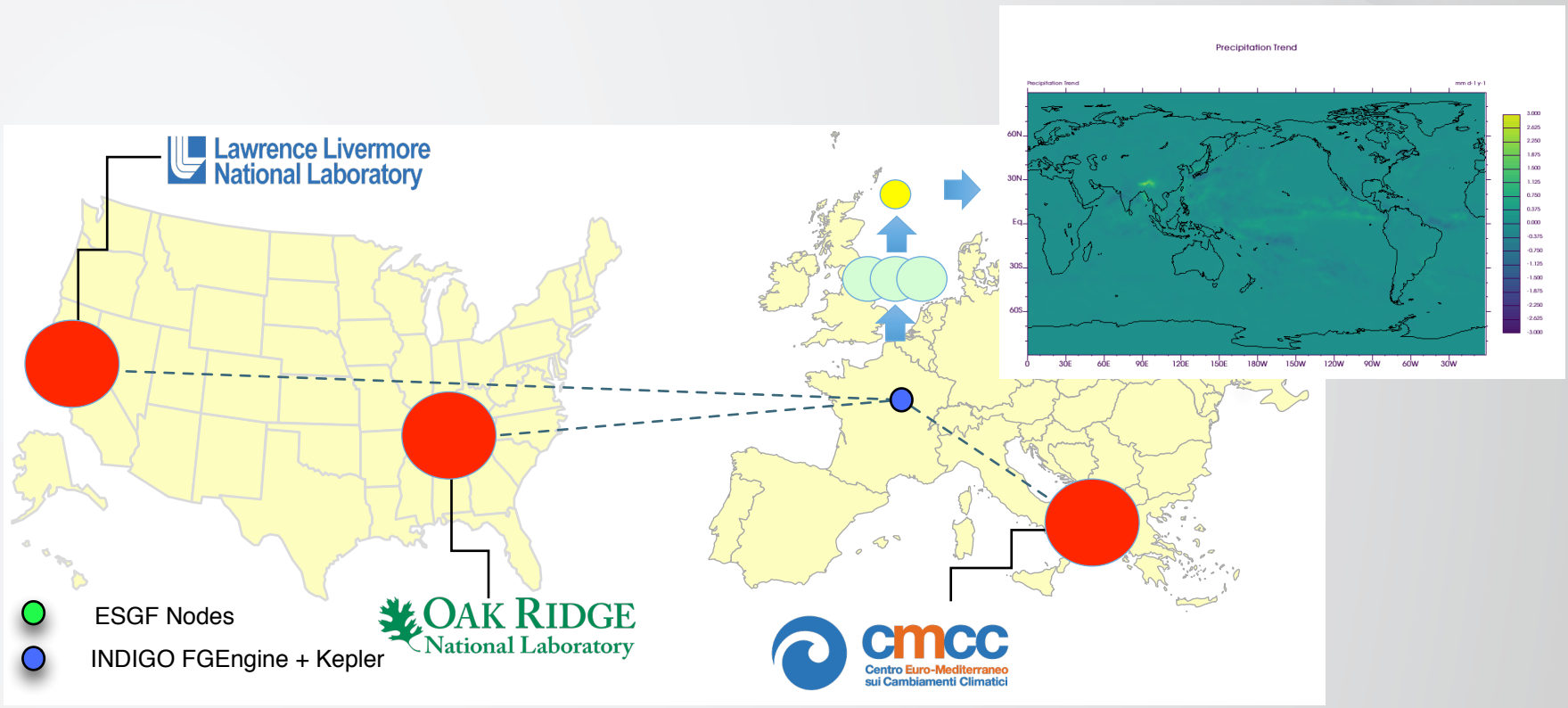
Single model precipitation trend analysis



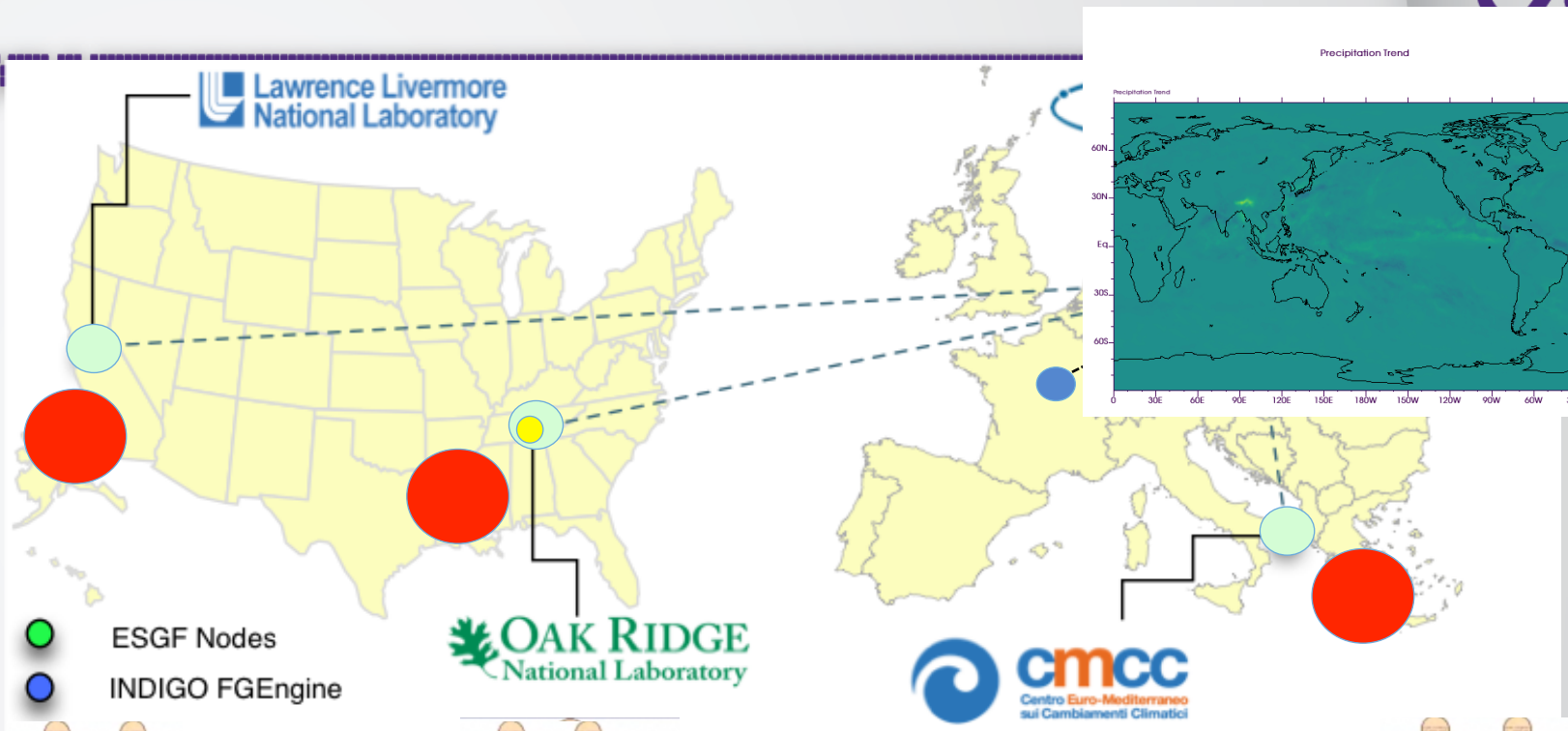
Climate Model Intercomparison Data Analysis case study challenges & issues

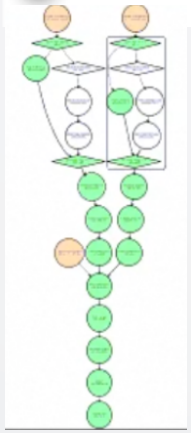
- CMIP* experiments provide input for multi-model analytics experiments (e.g. trend analysis)
  - Input data from multiple models needed
  - **Data distribution** inherent in the infrastructure
  - **Data download is a big barrier** for end-users (download can take from several days to weeks!)
  - Current infrastructure mainly for **data sharing**
  - **Data analysis** mainly performed using **client-side approaches**
  - Complexity of the data analysis needs more **robust end-to-end support**

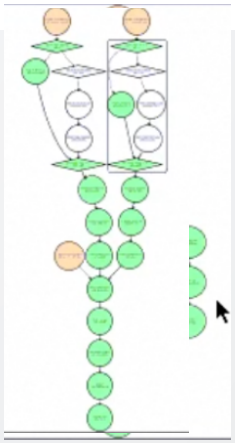
The current scientific workflow in ESGF (client-side)

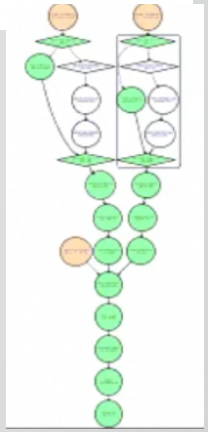


The paradigm shift implemented in INDIGO (server-side)





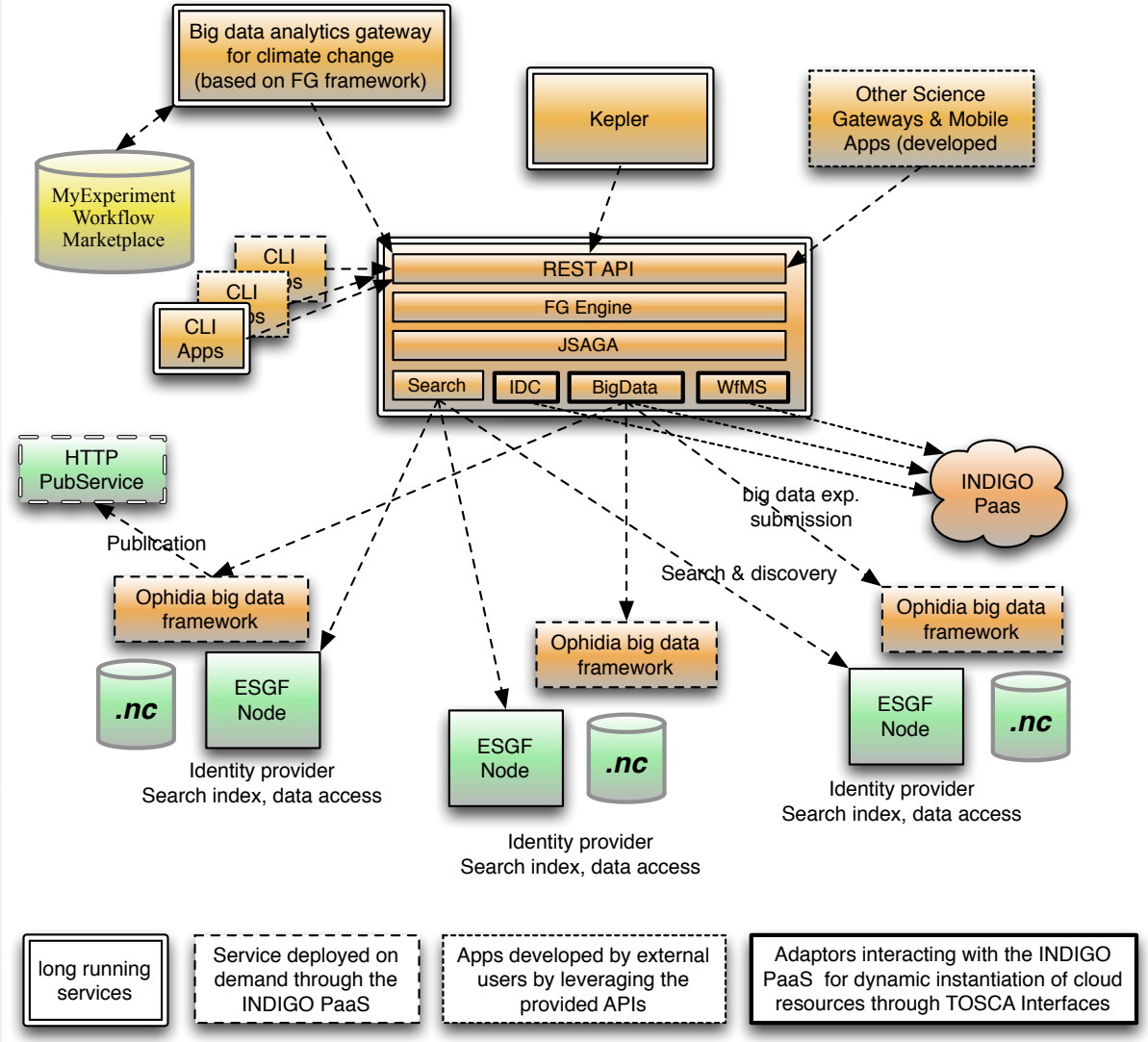




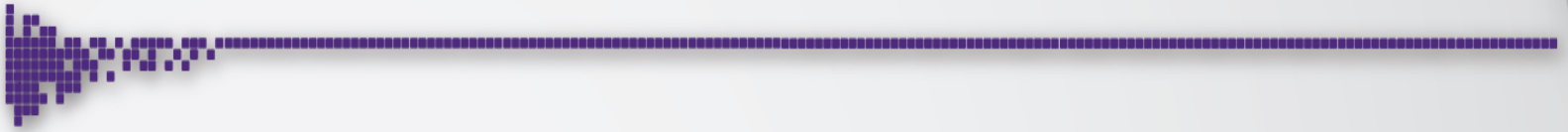
Architectural solution
Running the multi-model experiment

- Distributed experiments for climate data analysis
- Server-side processing
- Two-level workflow strategy to orchestrate multi-site experiments
- Three-level of parallelism
  - Inter-workflow, intra-workflow, intra-task
- Access through Kepler GUI
- INDIGO solutions: Kepler, FGEngine, Ophidia, INDIGO PaaS
- INDIGO complements, extends and interoperates with the ESGF stack

Legend: legacy components in green, INDIGO components in orange, external components in yellow



Running the multi-model experiment



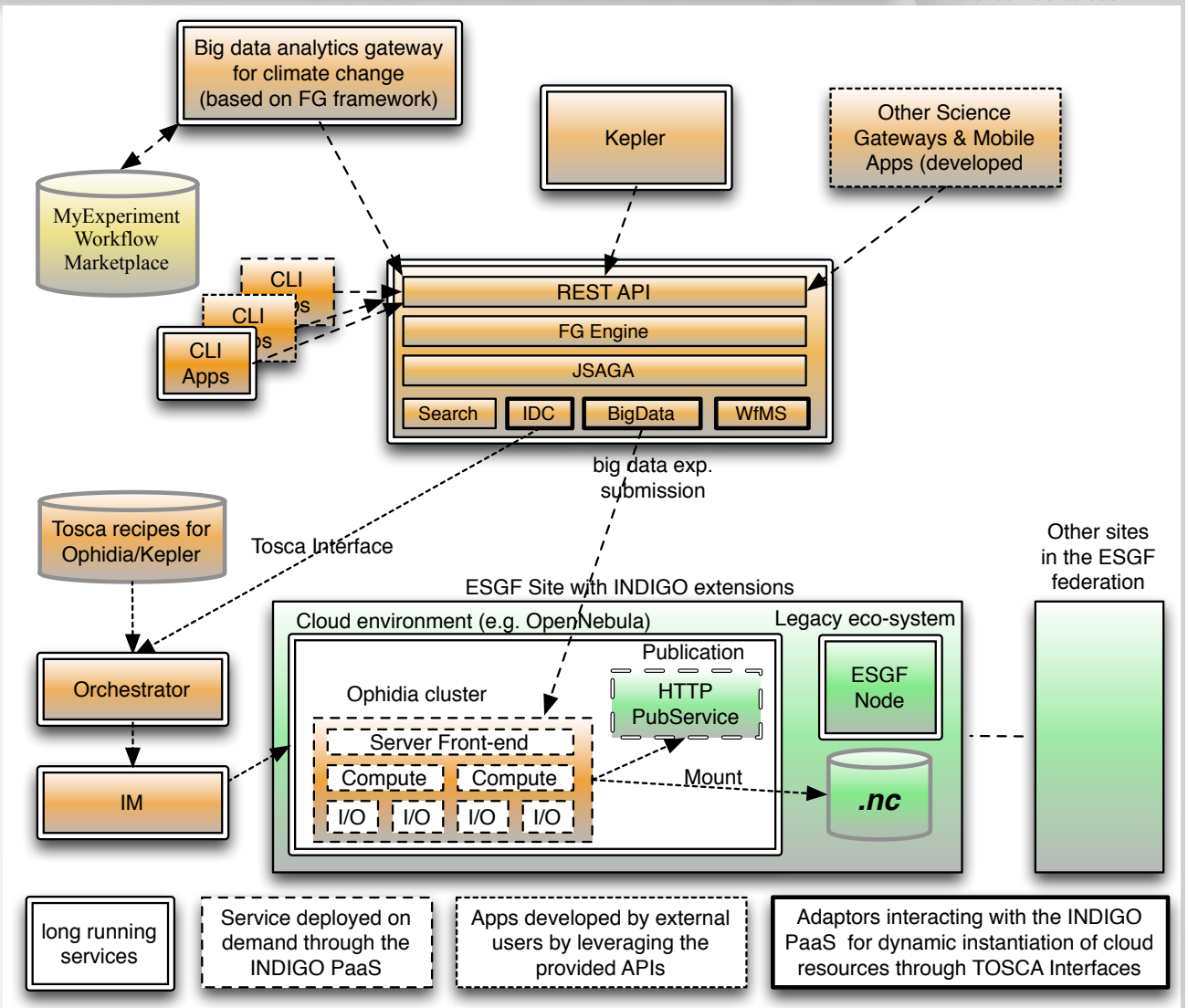
- **Application-domain oriented**
  - Strong requirements elicitation/validation
  - Prototype running on a **real testbed** involving **3 ESGF sites** + **PSNC**
  - **Integration of tools** widely used by the community (**UV-CDAT** data viz.)
  - Integrates multiple INDIGO components (**FGEngine, Kepler, Ophidia**)
    - Planned IAM, Orchestrator, CLUES, IM
  - Potential impact: **very high**
  - We expect the time-to-solution for the multi-model experiment can go down from **weeks** to **hours**!

Architectural solution
Flexible and dynamic deployment

INDIGO - DataCloud

- Dynamic instantiation of Ophidia and Kepler WfMS
- Automated deployment through TOSCA document
- Data locality key due to the large amount of data
- Interoperability with ESGF
- Integration of largely adopted community-based tools
  - UV-CDAT viz tool
  - OPeNDAP/THREDDS (publication services)

Legend: legacy components in green, INDIGO components in orange, external components in yellow



Flexible and dynamic deployment

- **Platform-as-a-Service level**
  - Dynamic deployment of Ophidia through the INDIGO PaaS layer
  - Based on ansible roles and TOSCA document
  - Run through the Command Line Interface

- **Dynamic and flexible deployment** of an Ophidia cluster
  - integrates multiple INDIGO components (**IAM, CLUES, IM, Orchestrator, Ophidia**)
  - **automates** and **makes easy** the deployment of an Ophidia cluster
    - Time-to-solution (deployment/setup) from 1-2 days to less than 1 hour!
  - enables the implementation of more “isolated” scenarios, where resources are deployed on demand on an experiment-basis

Added value and Innovation

INDIGO - DataCloud

Added Value

- **Paradigm shift** from **client-** to **server-side**
- Intrinsic **data movement reduction**
- Lightweight end-user setup
- **Re-usability** of data, final/intermediate products, workflows, etc.
- **Complements**, **extends** and **interoperates** with the **ESGF** stack
- Provisioning of a "**new** and **easy to use tool**" for scientists
- Drastic **time-to-solution reduction**

Innovation

- provisioning of a **core** infrastructural piece (based on **big data** and **cloud technologies**) enabling **large-scale data analysis** and **strongly needed** in the current **climate research ecosystem**

Exploitation: ESGF & RDA

- **Research Data Alliance**
  - Involvement into the **Array-Database Assessment WG**
  - **RDA application** with the aim of providing a *provenance-aware analytics eco-system* (ongoing evaluation – November 15, 2016)

- **Earth System Grid Federation**
  - Involvement into several ESGF Working Groups
  - Interaction with climate scientists from different ESGF sites
  - Testbed across EU/US involving 3 ESGF sites
  - Add new ESGF sites to the testbed

- **Goal: increase exploitation** and **users engagement!**

- **If you want to join the testbed, please contact us** (sandro.fiore@cmcc.it)

Dissemination events

- **EGU 2015** (12-17 April 2015, Vienna, Austria)
- **RDA** Sixth Plenary Meeting (23-25 September 2015, Paris, France)
- **EOScience2.0** (12-14 October 2015, Frascati, Italy)
- **ESGF F2F Conference** 2015 (7-11 December 2015, S. Francisco, CA, USA)
- **AGU2015** Conference (14-18 December 2015, S. Francisco, CA, USA)
- Ophidia PlayDay (29 April 2016, Bologna, Italy)
- Invited presentation at **LLNL** (23 May 2016, Livermore, CA, USA)
- Invited presentation at **ORNL** (26 May 2016, Oak Ridge, TN, USA)
- **CMCC Annual Meeting** (30-31 May 2016, Lecce)
- **Big Data and Extreme scale Computing** (15-17 June 2016, Frankfurt, Germany)
- **DI4R** (28-30 September 2016, Krakow, Poland)
- **ENES Community Meeting Reading 2016** (25-27 October 2016, Reading, UK)
- **ESGF F2F 2016 Conference** (Washington, December 6-9, 2016)

INDIGO - DataCloud

Thank you